Title: "The Genographic Project – a study of mtDNA and Y Chromosomes to Infer Human Migrations"

Gyan Bhanot: IBM Research and IAS Princeton

Abstract

The Genographic Project was launched in April 2005 jointly by IBM, the National Geographic Society and several population geneticists worldwide to collect DNA samples from indigenous peoples and the general public to study human migration and evolution from mtDNA and Y Chromosome markers.

The focus at IBM Research in the first year of the project has been on mtDNA. It is known that mtDNA suggest that all humans outside Africa originated from a founder population in Africa in one or more migrations approximately 50K-70K years ago, giving rise to the N clade found in Europe, Eurasia, North and South America and the M clade found in Central and South Asia, Australia and Oceania. However, the sequence, routes and precise timing of these migrations are still under debate. In this talk, I will present evidence from the analysis of 1737 complete public mtDNA sequences that shows that the M and N clades are genetically distinct and emerged from Africa in two separate migrations. The N clade split into two ancient and distinct groups which moved east and west. Contrary to current theory, this east-west split is represented on our tree not by the "R mega-group", but rather by a number of other mutations which replace the synonymous SNP at locus 12705 usually used to characterize R. Our analysis places the B, F and R5 haplogroups (which have a mutation at 12705 relative to L1) close to A/N9 (which do not have the mutation) and far from the other "R" haplogroups K/J/T/V/H. This is in agreement with the present geographic location of these groups (A, B, F, N9 and R5 mostly in China, Japan, South East Asia and the Americas and the J, T, H, U, and V mostly in Europe). Our results suggest that either SNP12705 had repeat mutations or appeared before the east-west split of N and was carried by both eastern and western migrating groups, becoming fixed in a heterogeneous way in the past 20K years. For the M clade, we identify a detailed substructure with 13 haplogroups of MD of which 6 are new. We also find a detailed substructure in the L0/L1, L2 and L3 clades. Distance distributions verify that L0/L1 is the oldest clade followed by L2 and L3/M/N. The L3, M and N clades are equidistant from each other and have the same amount of internal variation, which suggests that they are equally old in agreement with their origin from a common ancestor population.

Our method is a new technique which combines principal component analysis (PCA) and unsupervised consensus ensemble k-clustering (UCEkC) to identify robust clusters and ancient SNPs in all clades and haplogroups. It includes a method to directly create the tree of population events without the use of phylogeny techniques such as neighbor joining or parsimony. Our method also results in detailed assignment protocols for all robust haplogroups in the data with a high predictive accuracy. Our method is fast, can handle an arbitrary number of mutations and samples and mitigates the usual problems of sample size and sample choice bias. It was validated using extensive numerical simulations.